### Class Overview

<table>
<thead>
<tr>
<th>DNA Genomics</th>
<th>RNA Transcriptomics</th>
<th>Protein Proteomics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome wide comparison</td>
<td>Genome Comparison and Sequencing</td>
<td>Microarrays</td>
</tr>
<tr>
<td>Genome Comparison and Sequencing</td>
<td>Transcription factor binding sites</td>
<td>Orthology: Families of Proteins</td>
</tr>
<tr>
<td>Gene Finding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single sequence and Family comparison</td>
<td>Molecular Evolution</td>
<td>Experimental Techniques</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Secondary Structure Prediction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Molecular Evolution</td>
</tr>
</tbody>
</table>

**Tools:**
- String Alignment
- Phylogenetic Tree Construction
- Multiple Sequence AI
1. biological macromolecules - proteins and DNA/RNA
2. interaction of proteins, DNA and RNA is described by the Central Dogma of Molecular Biology - DNA is organized into genes, the genes are transcribed to RNA and translated to proteins
3. DNA, RNA
4. proteins
5. molecular evolution - continual process of random mutation and natural selection
Central Dogma

- DNA
  - Replication
  - Transcription
  - Reverse Transcription

- RNA
  - Translation

- Protein
Genome - The entire set of hereditary instructions for building, running and maintaining an organism. The genome encompasses the material passed on from generation to generation. (does not vary within one organism)

Genomics - study of the structure and function of the genome

Transcriptome - The entire set of messenger RNA expressed while building, running and maintaining an organism. The transcriptome is all the mRNA transcribed from genes within a given genome. (varies depending on physiology)

Transcriptomics - The genome-wide study of mRNA expression levels.

Proteome - The entire set of proteins expressed while building, running and maintaining an organism. The proteome is all the protein translated from mRNA of a given transcriptome. (varies with physiology)

Proteomics - study of the structure and function of the proteome
Biology

- Biology = Science of living things
- What is Life?
  - As a rule: living organisms exchange matter and energy with their environment, grow, develop, die, react to their environment and reproduce
  - Exceptions: seeds in vegetative state, viruses, viroids, prions...
  - Viruses
    - have DNA or RNA (genetic material)
    - evolve
    - can not use the genetic material themselves, they must have a host cell to function and reproduce
    - does not use enery themselves but the cells they infect do use energy
    - are they alive?
Computational Biology

- truly interdisciplinary
- biology is described by general rules with many many exceptions
- computer science is more clearly defined
- probability and statistics is important to describe general rules with many exceptions
- input data can be flawed
Composition of Organisms

- All organisms are based on similar biochemical processes.
- Main actors in these processes:
  - Proteins
  - Nucleic Acids (DNA/RNA)
DNA Vocabulary

nucleotide or nucleic acid or base one unit of DNA comprised of a base, a sugar and a phosphate group
oligonucleotide polymer of more than 1 nucleic acid
purine bases adenine and guanine
pyrimidine bases thymine and cytosine
transition conversion between a purine and a purine or a pyrimidine and a pyrimidine
transversion conversion between a purine and a pyrimidine
codon 3 bases that code for 1 amino acid
reading frame one of 3 possible ways to read a nucleotide sequences in one direction
frameshift type of mutation in which an insertion or deletion changes the reading frame
introns part of a gene that is transcribed but not present in a mature mRNA after splicing
exons part of a gene that is present in the mature mRNA and is translated to protein
Nucleic Acids

- Proteins are synthesized from information encoded in genes in DNA.
- **Nucleic Acids (NA)**
- 2 types of NA:
  1. Deoxiribonucleic acid (DNA)
  2. Ribonucleic acid (RNA)
DNA

- DNA consists of 2 strands
- Each strand has 3 components:
  1. Phosphate residue
  2. Sugar: 2’-deoxyribose
  3. 4 bases: Adenine (A), Cytosine (C), Guanine (G), Thymine (T)
Bases in DNA

pyrimidines Cytosine, Thymine
purines Adenine, Guanine
DNA

A series of 3 base pairs forms a "codon"

The codon is like a "word" in the "sentence" that makes a gene
RNA

The same as DNA except the sugar is ribose and the base Uracil (U) replaces Thymine (T):
RNA II

- single stranded → folds on to itself
- can take many different configurations
- has many different functions. 2 most important:
  1. mRNA: messenger RNA, information transport
  2. tRNA: transfer RNA
  3. rRNA: ribosomal RNA
  4. ribozymes: RNA with enzymatic functions
DNA Repair Mechanisms

Thymine dimer distorts the DNA molecule

Endonuclease cuts out the region containing the thymine dimer

Repair synthesis by DNA polymerase
Take Home Message: DNA/RNA

- DNA/RNA - strands have a repetitive pattern:
  - $[Sugar - Phosphate]$ –

  $\rightarrow$ chain structure

  $\rightarrow$ Orientation: 5’ $\rightarrow$ 3’

  $\rightarrow$ small number of bases and double strandedness of DNA insures secure data storage
Take Home Message: DNA/RNA II

Definition of alphabets:

\[ \mathcal{A}_{DNA} := \{A, C, G, T\} \quad \mathcal{A}_{RNA} := \{A, C, G, U\} \]

DNA sequence:

\[ s_{DNA} = B_1 \cdot B_2 \cdot \ldots \cdot B_n = \prod_{i=1}^{n} B_i \quad \text{where} \quad \forall_{i=1}^{n} B_i \in \mathcal{A}_{DNA} \]

RNA sequence:

\[ s_{RNA} = B_1 \cdot B_2 \cdot \ldots \cdot B_n = \prod_{i=1}^{n} B_i \quad \text{where} \quad \forall_{i=1}^{n} B_i \in \mathcal{A}_{RNA} \]
Proteins

- Main component of all organisms
- Perform most functions in the Cell
  1. **Structural proteins**: e.g. muscle filaments in a body
  2. **Enzymes**: catalysts for chemical reactions. Speed up of reaction and decrease in energy requirement.
  3. **Signalling Proteins**: e.g. cell receptors for estrogen)

- Proteins are built of simpler building blocks called:
  - **Amino Acids (AA)**
Amino Acids

Common structure of Amino Acids:

![Amino Acid Structure](image-url)
Amino Acids II

- In nature, 20 amino acids are common.
- Amino acids differ in their side chains (R).
- Side chains can be as simple as a hydrogen atom or as complex as two carbon rings.
- Characteristics of side chain that can influence their chemistry are: hydrophobicity, charge, acidity, basicity, and size.
Amino Acids II

[Diagram of amino acids]

- Alanine (A)
- Valine (V)
- Leucine (L)
- Isoleucine (I)
- Proline (P)
- Methionine (M)
- Phenylalanine (F)
- Tryptophan (W)
- Threonine (T)
- Cysteine (C)
- Asparagine (N)
- Glutamine (Q)
- Tyrosine (Y)
- Glutamic Acid (E)
- Aspartic Acid (D)
- Lysine (K)
- Arginine (R)
- Histidine (H)
Alle mini hydrophiles by Gina Cannarozzi

Alle mini hydrophiles
KREND, KREND
FAMILYVW are hydrophobes.
They dont like the sea.
“My name is Bond, Peptide Bond”

Question: How do amino acids form proteins? [→]

Peptide bonds
The central dogma of biology states that the coded genetic information hard-wired into DNA is transcribed into individual transportable cassettes, composed of messenger RNA (mRNA); each mRNA cassette contains the program for synthesis of a particular protein (or small number of proteins).
Transcription

Research questions:
Where are the introns? Where are the coding sequences? Where are the stop and start of translation? Where are the transcription factors that control when transcription takes place?
Translation

How is a mRNA sequence mapped to a protein?

→ The genetic code
  ▶ Groups of 3 subsequent bases are translated into 1 amino acid
→ Codons
## The Genetic Code

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GGG</td>
<td>G</td>
<td>Gly</td>
<td>AGG</td>
<td>R</td>
<td>Arg</td>
<td>CGG</td>
</tr>
<tr>
<td>GGA</td>
<td>G</td>
<td>Gly</td>
<td>AGA</td>
<td>R</td>
<td>Arg</td>
<td>CGA</td>
</tr>
<tr>
<td>GGC</td>
<td>G</td>
<td>Gly</td>
<td>AGC</td>
<td>S</td>
<td>Ser</td>
<td>CGC</td>
</tr>
<tr>
<td>GGU</td>
<td>G</td>
<td>Gly</td>
<td>AGU</td>
<td>S</td>
<td>Ser</td>
<td>CGU</td>
</tr>
<tr>
<td>GAG</td>
<td>E</td>
<td>Glu</td>
<td>AAG</td>
<td>K</td>
<td>Lys</td>
<td>CAG</td>
</tr>
<tr>
<td>GAA</td>
<td>E</td>
<td>Glu</td>
<td>AAA</td>
<td>K</td>
<td>Lys</td>
<td>CAA</td>
</tr>
<tr>
<td>GAC</td>
<td>D</td>
<td>Asp</td>
<td>AAC</td>
<td>N</td>
<td>Asn</td>
<td>CAC</td>
</tr>
<tr>
<td>GAU</td>
<td>D</td>
<td>Asp</td>
<td>AAU</td>
<td>N</td>
<td>Asn</td>
<td>CAU</td>
</tr>
<tr>
<td>GCG</td>
<td>A</td>
<td>Ala</td>
<td>ACG</td>
<td>T</td>
<td>Thr</td>
<td>CCG</td>
</tr>
<tr>
<td>GCA</td>
<td>A</td>
<td>Ala</td>
<td>ACA</td>
<td>T</td>
<td>Thr</td>
<td>CCA</td>
</tr>
<tr>
<td>GCC</td>
<td>A</td>
<td>Ala</td>
<td>ACC</td>
<td>T</td>
<td>Thr</td>
<td>CCC</td>
</tr>
<tr>
<td>GCU</td>
<td>A</td>
<td>Ala</td>
<td>ACU</td>
<td>T</td>
<td>Thr</td>
<td>CUC</td>
</tr>
<tr>
<td>GUG</td>
<td>V</td>
<td>Val</td>
<td>AUG</td>
<td>M</td>
<td>Met</td>
<td>CUG</td>
</tr>
<tr>
<td>GUA</td>
<td>V</td>
<td>Val</td>
<td>AUA</td>
<td>I</td>
<td>Ile</td>
<td>CUA</td>
</tr>
<tr>
<td>GUC</td>
<td>V</td>
<td>Val</td>
<td>AUC</td>
<td>I</td>
<td>Ile</td>
<td>CUC</td>
</tr>
<tr>
<td>GUU</td>
<td>V</td>
<td>Val</td>
<td>AUU</td>
<td>I</td>
<td>Ile</td>
<td>CUU</td>
</tr>
</tbody>
</table>
Translation in the Ribosome
Formal representation of proteins

- Each amino acid represented by one letter
- all amino acids as alphabet:

- Together with chain structure and orientation
- proteins can be represented as **strings**

\[ p = a_1 \cdot a_2 \cdot ... \cdot a_n = \prod_{i=1}^{n} a_i \quad \text{where} \quad \forall_{i=1}^{n} a_i \in A \]

where “\cdot” denotes the concatenation operator (Diss Knecht, 1996)
Protein Structures

- From a biological point of view proteins are not just strings
- Proteins fold into 3D-configurations
- Each amino acid can behave differently depending on the situation
- Characterization of proteins with the following structures:
  - Primary: sequence
  - Secondary: alpha helix or beta strand which arrange into sheets
  - Tertiary: packing of secondary structures
  - Quaternary: higher order packing
Primary Structure

- the sequence of 20 amino acids
  KLKDYKHAYHPVDLDIKDIDYTMFHLACITKLF
- linear (unbranched)
- Repetitive motive of: \([- N - C - (CO) - ]\)
  \(\rightarrow\) backbone \(\rightarrow\) chain structure
- Orientation from \(N\)– to \(C\)– terminal end
- 50-3000 amino acids
Secondary Structure- alpha Helix (H)

- alpha helices have 3.6 residues per turn. (100 degrees separate the side chains of consecutive residues)

- there is a hydrogen bond between residue i and residue i+4

- average length of an alpha helix is 11 amino acids although they can be as short as 5 and as long as 25
Secondary Structure- beta strands
Tertiary Structure- the "fold" of the protein

Some protein folds occur in multiple genomes

- P-loop hydrolase
- Flavodoxin-like
- Rossman fold
- Thiamin binding
- TIM barrel
Summary

[Diagram of protein structure: primary structure, secondary structure (beta pleated sheet, alpha helix), tertiary structure, quaternary structure]
crystal structure of Triose Phosphate Isomerase

Dimer (two identical proteins together) with beta-barrel fold.
Protein Vocabulary

amino acid or residue building blocks of proteins
backbone the N-C-C -N-C-C bonds that form the chain of the protein
side chains the part of each amino acid that gives the distinctive characteristic and carries the functionality (acidic, basic, charged + or -, hydrophobic, hydrophilic)
fold refers to the 3 dimensional structure - common folds are, for example, a barrel or an alpha/beta sandwich
hydrophobicity characteristic of amino acid side chains that determines much of how they behave- are they compatible with water or not?
hydrophobic do not like water (FAMILYVW)
hydrophilic like water (KREND)
hydrogen bonding a strong type of intermolecular attraction based on partial charges. Occurs between hydrogen and F, O and N.
alpha helix a helix formed in a protein (not the same as double helix of DNA)- characterized by hydrogen bonds between residue i and i+4
beta strand an extended chain conformation in a protein in which the side chains stick out on alternating sides and hydrogen bonding is between neighboring strands in sheets. Beta strands are almost always arranged into sheets.
Molecular Evolution

- Over time DNA sequences change due to
  - UV-radiation, Toxins, Radioactivity, other factors ...
- ... as a consequence, DNA undergoes
  - Mutations: Change of one base to another base
  - Deletions of bases
  - Insertions of bases
- Other evolutionary forces
Evolution, Phylogeny and the Alignment

- alignment implies an evolutionary relationship represented by Phylogenetic Tree
- alignment shows accepted substitutions since divergence
- our model allows for point mutations and insertions/deletions (indels)
- aligns amino acids that diverged from the same residue in (hypothetical) most recent common ancestor
- darwinian evolution is driven by mutation and natural selection
- mutations may be adaptive, neutral or deleterious
- proteins evolve under functional constraints - mutations that destroy function do not appear in database via organism death
- "correct" alignment represents actual events- substitutions, indels
- impossible to verify -> take alignment with the highest probability that the alignment is correct under our model
Tree of Life

All genomes, Wed Aug 3 11:11:36 2005, 197 species, 132924 orthologs, 11 used here, distance tree, ToComple=38, Fit=0.4633; modified for educational purposes (dmm)

tree courtesy of Daniel Margadant
Molecular Evolution Vocabulary

**homology** related by common ancestry
**homologs** biological sequences related by common ancestry
**orthology** homology that arises via speciation
**paralogy** homology that arises via gene duplication
**eukaryote** one of the 3 kingdoms of life, organisms that have a nucleus
**prokaryote** organisms that lack a nucleus
**archaea** one of the 3 kingdoms of life, prokaryotes that often live in extreme environments (hot, salty, cold, basic, acidic)
**bacteria** one of the 3 kingdoms of life, prokaryotes
Web Sites

- Course:  
  http://www.icos.ethz.ch/education/courses/computationalbiology
- CBRG:  
  http://www.cbrg.ethz.ch/
- BioRecipes:  
  http://www.biorecipes.com